D152–D157 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 30 October 2010*
doi:10.1093/nar/gkq1027

miRBase: integrating microRNA annotation and deep-sequencing data

Ana Kozomara and Sam Griffiths-Jones*

Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

Received September 15, 2010; Accepted October 10, 2010

ABSTRACT

miRBase is the primary online repository for all microRNA sequences and annotation. The current release (miRBase 16) contains over 15 000 microRNA gene loci in over 140 species, and over 17 000 distinct mature microRNA sequences. Deep-sequencing technologies have delivered a sharp rise in the rate of novel microRNA discovery. We have mapped reads from short RNA deep-sequencing experiments to microRNAs in miRBase and developed web interfaces to view these mappings. The user can view all read data associated with a given microRNA annotation, filter reads by experiment and count, and search for microRNAs by tissue- and stage-specific expression. These data can be used as a proxy for relative expression levels of microRNA sequences, provide detailed evidence for microRNA annotations and alternative isoforms of mature microRNAs, and allow us to revisit previous annotations. miRBase is available online at: http://www.mirbase.org/.

INTRODUCTION

miRBase is the primary online repository for microRNA sequences and annotations. The main aims of miRBase are:

(1) to curate a consistent nomenclature scheme by which novel microRNAs are named;
(2) to act as the central repository for all published microRNA sequences, and to facilitate online searching and bulk download of all microRNA data;
(3) to provide human-readable and computer-parsable annotation of microRNA sequences (for example, functional data, references, genome mappings);
(4) to provide access to the primary evidence that supports microRNA annotations; and
(5) to link to and aggregate microRNA target predictions and validations.

The miRBase database was established in 2002 (then called the microRNA Registry) to provide microRNA researchers with stable and unique gene names for their novel microRNA discoveries (aim 1) and an archive of all microRNA sequences (aim 2) (1–3). Official gene names assigned by miRBase should be used in the published version of articles that describe their identification; gene names are assigned in confidence for inpress manuscripts (1). The incorporation of short RNA deep-sequencing data into miRBase as evidence for microRNA annotation (aim 4) is described in this update. Expansion of textual and functional annotation (aim 3), and an aggregation service for the growing number of microRNA target predictions and validations (aim 5) are the subject of future work.

From its inception, miRBase was designed to be a focused resource that could make and facilitate significant contributions in a rapidly growing field. We currently capture data types from user submissions and from publications that describe novel microRNAs, for example, the experimental method used to identify the sequence. Alongside the microRNA name, miRBase assigns a stable accession number to each stem–loop and mature sequence to allow tracking of improved annotations between releases. The primary transcripts of almost all microRNAs remain unannotated, but we aim to develop a mechanism to include annotations as they are determined. Where genome assemblies are available, microRNAs are mapped to their locations, clusters of microRNAs are highlighted, and overlaps with annotated protein-coding genes are described. Families of microRNAs are constructed, and links are provided to entries in other databases, to predicted targets, and to the primary literature. miRBase provides several methods to access the sequence data: by browsing, sequence similarity, genomic coordinate intervals, keyword search and bulk download. MicroRNA gene nomenclature and details of the data types available in

*To whom correspondence should be addressed. Tel: +44 161 2755673; Fax: +44 161 2755082; Email: sam.griffiths-jones@manchester.ac.uk

miRBase have been discussed previously (1–3). We focus here on growth of the database, recent developments to integrate deep-sequencing data, and future plans.

DATABASE GROWTH

miRBase is a vital tool for microRNA research. In order to maintain the usefulness of the resource, we must develop tools keep pace with increased rates of microRNA discovery that have been facilitated by next-generation technologies such as deep sequencing. The number of microRNAs deposited in miRBase has risen approximately exponentially (Figure 1). In the last 3 years alone, the number of microRNA sequences in the database has almost trebled. At the time of writing, miRBase (release 16) contains over 15 000 microRNA loci, expressing over 17 000 distinct mature sequences, from 142 species.

Current semi-automated procedures for building miRBase entries from submitted data and from supplementary data to publications have been sufficient to keep pace with historical rates of microRNA identification. Post-2007, almost all of the growth of miRBase has been driven by deep-sequencing experiments. Each experiment may discover 10s or 100s of novel microRNAs. Many novel sequences are specific to a studied tissue or stage and are not conserved between species. Each experiment may also provide evidence for a large number of pre-existing miRBase entries. The computational and curational challenge involved in dealing with the increased data volume is significant. Our initial developments and improvements to incorporate data from deep-sequencing experiments into miRBase involve a sequence and read visualization tool and a pipeline for rapid mapping. These improvements provide the infrastructure to track increased rates of sequence deposition in the coming years. However, the number of publications that describe microRNA experiments and functional analysis is also increasing exponentially (Figure 1). For example, in 2009 alone, over 2300 articles in PubMed reference the keyword 'microRNA' (including 477 reviews), only 8 years after the first use of the term in the literature. Curating the functional annotation contained in this corpus is currently impossible. Recent improvements in text-mining approaches for biomedical literature [reviewed in refs (4,5)] provide some hope that automated textual annotation of microRNAs may become feasible in the medium term.

INCORPORATING DATA FROM DEEP-SEQUENCING EXPERIMENTS

As discussed, the majority of the evidence supporting microRNA annotations now comes from deep-sequencing experiments. We have developed an interface to view reads from RNA deep-sequencing data mapped to microRNA loci on the miRBase web site. Briefly, we identify entries in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with short-read data that map to references discovering novel microRNAs. We extract the read sequences and counts from the GEO entry and map the reads to the set of miRBase hairpin sequences for a given organism using Bowtie (6) allowing at most two mismatches between the read and the hairpin sequence.

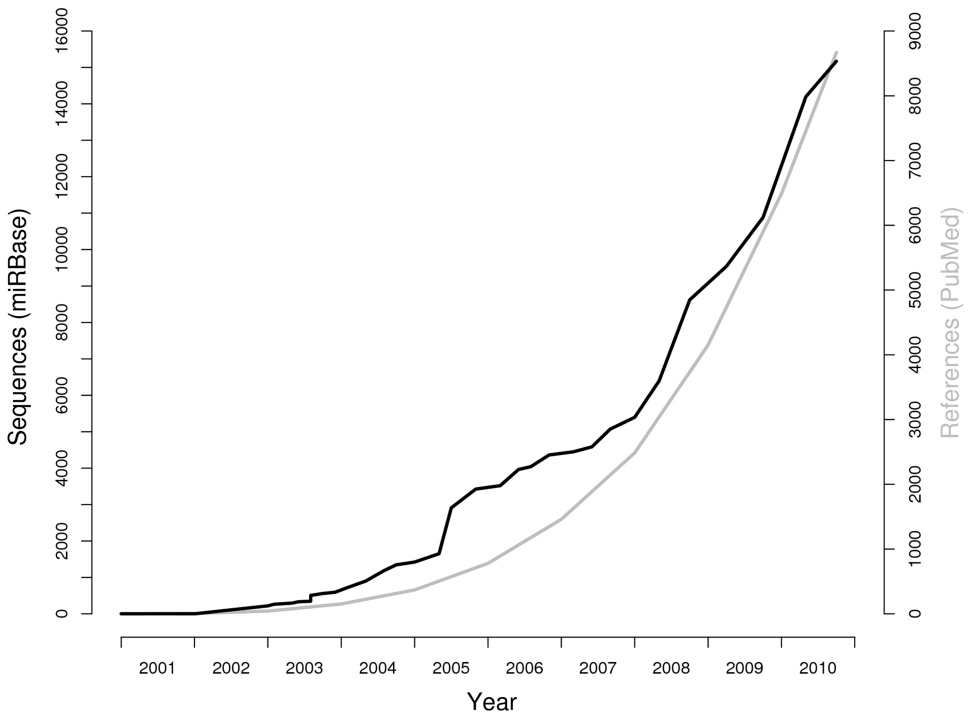

Figure 1. Growth of the miRBase sequence archive (black) and the microRNA publication record (PubMed entries that reference the term 'microRNA', grey).

The first deep-sequencing data sets incorporated into miRBase are from human, *Drosophila melanogaster*, *Arabidopsis thaliana*, rice and three nematode genomes. Additional data sets will be added rapidly. Ideally, all microRNA sequences submitted to miRBase will be linked to GEO submissions, or followed by the submission of the deep-sequencing data for inclusion in miRBase.

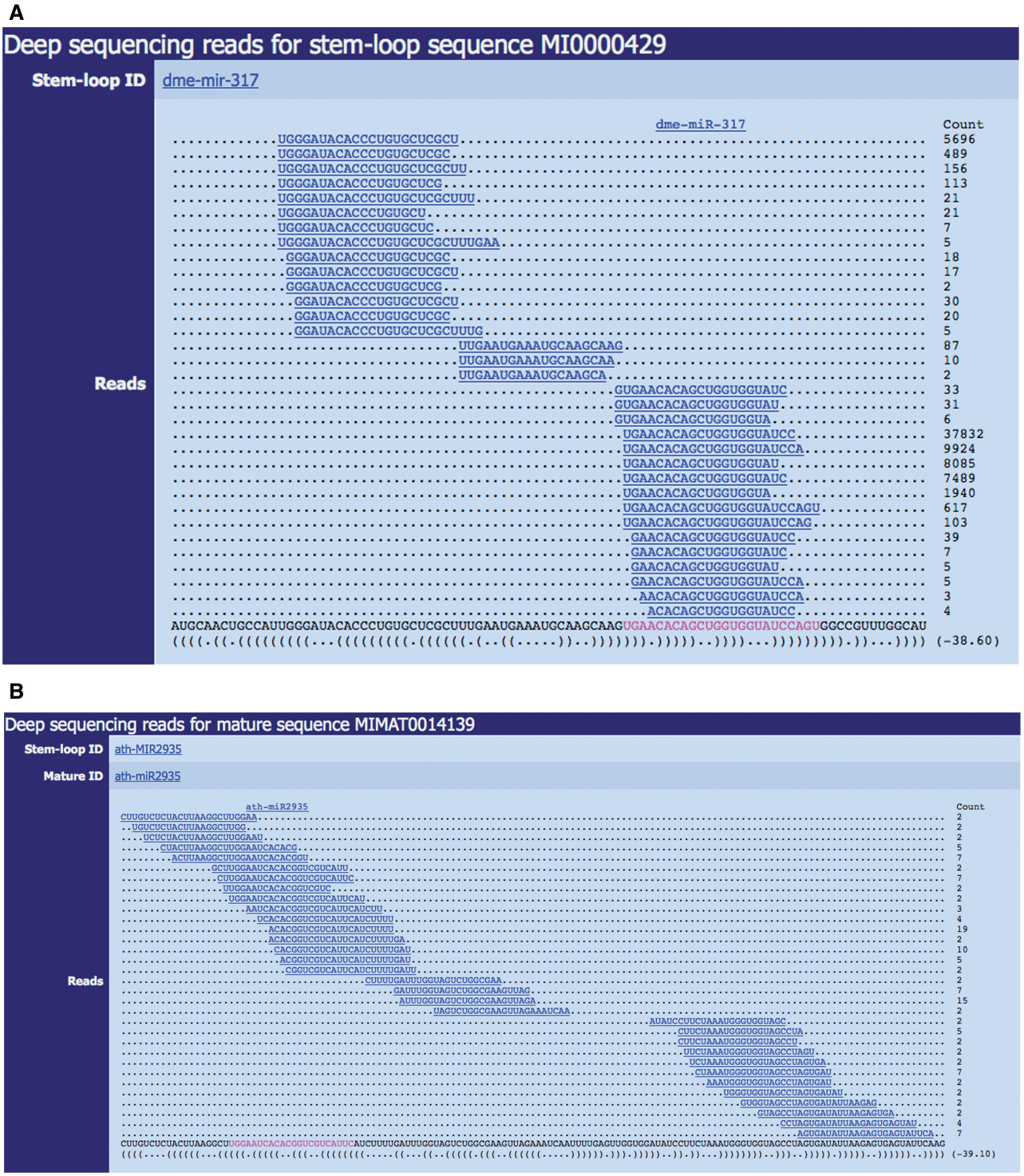

Figure 2. miRBase view of deep-sequencing reads mapping to (**A**) dme-mir-317 and (**B**) ath-MIR2935 loci. Each unique read is mapped to the stem-loop sequence (bottom), with the annotated mature sequence in miRBase highlighted (magenta), and the sum aggregated read counts from all stored experiments shown (right). Reads with only a single copy are hidden. Data shown are from 14 experiments [dme-mir-317 (15)] and four experiments [ath-MIR2935 (16)].

The view of deep-sequencing data can be accessed from miRBase stem–loop and mature pages (Figure 2). Reads can be filtered by the number of mismatches to the hairpin sequence, the read count and by experiment. Each experiment is annotated with species, tissue, stage and methodology information. These tags enable the user to search for experiments by tissue expression on the miRBase search page (http://www.mirbase.org/search.shtml). Read counts for mature microRNAs are commonly used as a proxy for relative expression levels. As the number of deep-sequencing experiments increases, these data provide extensive information about the expression profiles of microRNAs across tissues, stages and organisms. A good example of the power of these data can be seen in the small RNA data made available through the MODENCODE project for *D. melanogaster* (7). The patterns of read mappings also provide valuable evidence for relative abundance of mature sequences from different arms, for isoforms of mature microRNA sequences, and for the confidence in a given microRNA annotation. For example, Figure 2 shows a number of isoforms of mature microRNAs derived from both arms of the dme-mir-317 hairpin precursor. The dominant mature sequence from the 3′-arm is 21 nt in length, whereas the mature sequence annotated in miRBase 16 is extended to 24 nt. The 5′-mature sequence is unannotated in miRBase 16. The miRBase view of read data therefore provides the user with a clear picture of the profile of short RNAs generated from a microRNA locus, and facilitates the correction of errors and omissions in future releases.

DEEP-SEQUENCING DATA AIDS DISCRIMINATION BETWEEN MICRORNAS AND OTHER RNA SPECIES AND FRAGMENTS

Guidelines for microRNA annotation were established in 2003 (8), requiring evidence of expression of a ~22 nt sequence (for example, cloning, sequencing or northern blot), together with evidence for a microRNA precursor structure (predicted stem–loop flanking the mature sequence). Updated annotation criteria were recently suggested to distinguish microRNAs from other classes of short RNAs in plants (9). These standards have proved extremely powerful in maintaining a clean data set of microRNA sequences for the community.

The increased rates of microRNA detection afforded by deep-sequencing technologies provide challenges to the level of confidence required to annotate a sequence as a microRNA. A typical RNA deep-sequencing experiment will identify millions of short sequences. Increased coverage results in detection of sequences of ever-lower abundance. It therefore becomes more and more challenging to distinguish true microRNAs from fragments of other transcripts, other short RNAs and spurious transcription. The eukaryotic genome also contains millions of predicted hairpins, so a flanking stem–loop structure should be considered necessary but not sufficient to annotate a sequence as a microRNA. If poorly analysed, a single data set thus has the potential to generate a large number of dubious annotations, swamping the real microRNAs. However, correct interpretation of RNA deep-sequencing data provides several additional signals to help distinguish microRNAs from other sequences. A number of recent publications have attempted to define and use criteria based on patterns of mapped reads (10–14), and a consensus set of guidelines is starting to emerge:

(1) Multiple reads (10–20 are commonly used cutoffs) support the presence of the mature microRNA (preferably from multiple independent experiments);
(2) The reads map to an extended sequence region (e.g. an assembled contig), and the sequence flanking the putative mature microRNA folds to form a microRNA precursor-like hairpin with strong pairing between the mature microRNA and the opposite arm. Reads that map very many times to a genome sequence should be discarded;
(3) Mapped reads do not overlap other annotated transcripts (i.e. there is no evidence that the short reads may represent fragments of mRNAs or other known RNA types);
(4) Reads mapping to a locus support consistent processing of the 5′-end of the mature sequence (for example, the majority of reads overlapping a given mature microRNA annotation should have the same 5′-end; the 3′-end may be significantly more variable); and
(5) Ideally, reads will support the presence of mature sequences from both arms of the predicted hairpin (so-called miR and miR* sequences), and the putative mature sequences should base-pair with the correct 3′-overhang.

Consistent 5′-end processing (point 4 above), and observation of miR and miR* sequences (point 5) appear to be crucial for discrimination between high-confidence microRNAs and fragments of other RNAs in deep-sequencing data (10–12). Figure 2A shows the miRBase view of deep-sequencing reads mapping to the dme-mir-317 precursor region. The pattern of reads clearly supports a high-confidence microRNA annotation, with over 65 000 reads from 14 experiments (15) supporting the 5′-end of a mature sequence derived from the 3′-arm of the hairpin, and over 5000 reads supporting the miR* sequence [unannotated in miRBase (16)]. In contrast, the pattern of reads overlapping the ath-MIR2935 sequence does not support the annotation of a microRNA (Figure 2B), with multiple offset reads distributed across the locus. In addition, the *Arabidopsis* reads shown are isolated from different Argonaute complexes—the majority of reads are not associated with the microRNA AGO1 complex, rather with AGO4 (16). These data suggest that MIR2935 should be removed from the miRBase microRNA catalogue in future releases.

FUTURE DEVELOPMENTS

There are three main areas of future development of miRBase currently planned. We invite feedback and comments on these areas.

Improvement of community contribution of and access to microRNA data

miRBase is a community resource. The majority of the primary sequence data is submitted by users. We plan to improve methods and interfaces for both data access and data submission. For example, webservices will allow programmatic access to all miRBase data, and batch search and download tools will be made available. We plan to allow users to add and update textual annotation in a wiki interface. The Rfam database of RNA families (17) has successfully developed a community annotation project using Wikipedia (18).

Expansion to include microRNA candidates and predictions

As discussed above, deep-sequencing experiments provide the majority of novel microRNA annotations. Next-generation experimental techniques also provide low-level evidence for comparatively large numbers of lower confidence sequences, often published as candidate microRNAs, which fall below the current standards and are therefore not present in miRBase. We plan to extend the database to include different classes of data as supplements, with associated confidence levels for each annotation, for example:

(1) High-confidence microRNA genes, with support for the presence of both miR and miR*, sequenced in many copies from many independent samples, as described above;
(2) Predicted homologues of known microRNAs in related organisms;
(3) Low abundance sequences, cloned or sequenced in only a few copies, without miR* evidence;
(4) Sequenced short RNAs that are absent from assembled genomes; and
(5) Computational microRNA predictions.

High-confidence annotations will form the default microRNA set that is most relevant to the majority of users. Lower confidence annotations will be provided as supplements to those data. Their inclusion will encourage validation of low confidence sequences, and the availability of experimental resources (for example, off-the-shelf microarrays).

A microRNA target aggregation service

From 2006, miRBase provided a set of microRNA target predictions, branded miRBase targets, in collaboration with the Enright lab (2). In 2009, the miRBase targets resource was devolved back into the Enright lab at the EBI, and rebranded microCosm. In its place, we will develop an aggregation service that integrates microRNA target predictions from all the popular target prediction sites [for example, TargetScan (19), PicTar (20), microCosm, DIANA-microT (21), microrna.org (22)], together with lists of validated microRNA target sites [currently curated by resources such as TarBase (23), miRecords (24), miRTarBase]. This will initially involve displaying the top hits from each algorithm on existing miRBase sequence pages. In the longer term, an interface to search for hits from the aggregated algorithms will be developed, and consensus predictions from subsets of the algorithms will be provided. All results will link to the original data sources. To eliminate any long-term curation burden, we will accept depositions of microRNA target predictions and validated targets from users and target prediction groups.

AVAILABILITY

The miRBase database is available online and free to all without restriction at: http://www.mirbase.org/ and for download in various formats (including FASTA sequences, GFF genome coordinates and MySQL database dumps) from ftp://mirbase.org/pub/mirbase/CURRENT/. Nomenclature queries, feedback and comments are welcomed at mirbase@manchester.ac.uk.

ACKNOWLEDGEMENTS

We thank the Wellcome Trust Sanger Institute for previous hosting and support, Simon Moxon for insight into deep-sequencing data in plants and Antonio Marco and Matthew Ronshaugen for helpful comments on the article.

FUNDING

miRBase is funded by the Biotechnology and Biological Sciences Research Council (BB/G022623/1). Funding for open access charge: The BBSRC (BB/G022623/1).

Conflict of interest statement. None declared.

REFERENCES

1. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.
2. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
3. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
4. Jensen,L.J., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
5. Krallinger,M., Valencia,A. and Hirschman,L. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9**, S8.
6. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
7. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
8. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
9. Meyers,B.C., Axtell,M.J., Bartel,B., Bartel,D.P., Baulcombe,D., Bowman,J.L., Cao,X., Carrington,J.C., Chen,X., Green,P.J. *et al.*

(2008) Criteria for annotation of plant MicroRNAs. *Plant Cell*, **20**, 3186–3190.
10. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al*. (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev*., **24**, 992–1009.
11. Berezikov,E., Liu,N., Flynt,A.S., Hodges,E., Rooks,M., Hannon,G.J. and Lai,E.C. (2010) Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila*. *Nat. Genet*., **42**, 6–9.
12. Marco,A., Hui,J.H., Ronshaugen,M. and Griffiths-Jones,S. (2010) Functional shifts in insect microRNA evolution. *Genome Biol. Evol*., **2**, 686–696.
13. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol*., **26**, 407–415.
14. Hendrix,D., Levine,M. and Shi,W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol*., **11**, R39.
15. Chung,W.J., Okamura,K., Martin,R. and Lai,E.C. (2008) Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. *Curr. Biol*., **18**, 795–802.
16. Mi,S., Cai,T., Hu,Y., Chen,Y., Hodges,E., Ni,F., Wu,L., Li,S., Zhou,H., Long,C. *et al*. (2008) Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. *Cell*, **133**, 116–127.
17. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al*. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res*., **37**, D136–D140.
18. Daub,J., Gardner,P.P., Tate,J., Ramskold,D., Manske,M., Scott,W.G., Weinberg,Z., Griffiths-Jones,S. and Bateman,A. (2008) The RNA WikiProject: community annotation of RNA families. *RNA*, **14**, 2462–2464.
19. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*., **19**, 92–105.
20. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet*., **37**, 495–500.
21. Maragkakis,M., Reczko,M., Simossis,V.A., Alexiou,P., Papadopoulos,G.L., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K. *et al*. (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*., **37**, W273–W276.
22. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res*., **36**, D149–D153.
23. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*., **37**, D155–D158.
24. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*., **37**, D105–D110.